DERIVATION OF RISSANEN'S "MDL DENOISING" CRITERION

DANIEL F. SCHMIDT

Abstract. This note derives Rissanen's "MDL Denoising" criterion for linear regression models based on the normalized maximum likelihood density in an expanded form.

1. Introduction

This document discusses and explains the derivation of the normalized maximum likelihood model selection criterion for linear regression models presented by J. Rissanen in [1]. In the setting of a Gaussian linear regression model, the target y_j is assumed to be normally distributed with variance τ and conditional mean equal to the weighted linear combination of k associated covariates $\bar{\mathbf{x}}_j = (x_{j,1}, \ldots, x_{j,k})'$, the combination weightings being determined by the regression parameters, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$; that is,

$$y_j | \bar{\mathbf{x}}_j, \boldsymbol{\beta} \sim N(\bar{\mathbf{x}}_j \boldsymbol{\beta}, \tau). \tag{1.1}$$

The likelihood of a data vector $\mathbf{y} = (y_1, \ldots, y_n)$, conditional on the $(n \times k)$ regressor matrix $\mathbf{X} = (\bar{\mathbf{x}}_1', \ldots, \bar{\mathbf{x}}_n')'$, regression coefficients $\boldsymbol{\beta}$, and noise variance parameter τ is

$$p(\mathbf{y}; \boldsymbol{\beta}, \tau) = \left(\frac{1}{2\pi\tau} \right)^{\frac{n}{2}} \exp\left(-\frac{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2}{2\tau} \right). \tag{1.2}$$

Let $\boldsymbol{\Sigma} = \mathbf{X}'\mathbf{X}$; the maximum likelihood estimates are given by (see the Appendix for details)

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\mathbf{y}) &= \boldsymbol{\Sigma}^{-1}\mathbf{X}'\mathbf{y} \qquad (1.3) \\ \hat{\tau}(\mathbf{y}) &= \left(\frac{1}{n} \right) (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y}))'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\mathbf{y})) \\ &= \left(\frac{1}{n} \right) \mathbf{y}' \left(\boldsymbol{I}_n - \mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{X} \right) \mathbf{y}. \qquad (1.4) \end{aligned}$$

A standard problem is to determine which, if any, of the covariates is associated with the target vector $\mathbf{y}$. There is an extremely large body of literature studying this problem, and Rissanen [1] derived an information theoretic covariate selection criterion based on the normalized maximum likelihood (NML) [2] density given by

$$\bar{p}(\mathbf{y}) = \frac{p(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}))}{\int_{\mathbf{z} \in Y} p(\mathbf{z}; \hat{\boldsymbol{\beta}}(\mathbf{z}), \hat{\tau}(\mathbf{z})) d\mathbf{z}}, \tag{1.5}$$

where $\hat{\boldsymbol{\beta}}(\mathbf{y})$ and $\hat{\tau}(\mathbf{y})$ denote the maximum likelihood estimates of $\boldsymbol{\beta}$ and τ respectively. The negative logarithm of the normalized maximum likelihood (NML) density may be interpreted as a measure of model fit to the data, and in this capacity can be used to determine whether a covariate should be included in the model. The more covariates that are included, the better the fit of the linear model, and thus the greater the maximised likelihood, as measured by the numerator of (1.5). However, the denominator of the NML density acts to prevent "overfitting" by penalizing models that contain a greater number of covariates. The model that minimises the NML score is considered the "optimal" model.

2. First-Level Universal Model

For the linear-Gaussian regression model, the numerator in (1.5) is simply given by

$$p(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y})) = \left(\frac{1}{2\pi e \hat{\tau}(\mathbf{y})} \right)^{\frac{n}{2}}. \tag{2.1}$$

It remains to evaluate the denominator in (1.5). Unfortunately, this diverges if the region of integration Y is taken to be the entirety of the n-dimensional dataspace. To avoid this problem, Rissanen chose to integrate over a subset of $\mathbb{R}^n$ defined by

$$Y_1(R, \tau_0) = \left\{ \mathbf{y} \in \mathbb{R}^n : \hat{\boldsymbol{\beta}}'(\mathbf{y}) \boldsymbol{\Sigma} \hat{\boldsymbol{\beta}}'(\mathbf{y}) \leq nR, \hat{\tau}(\mathbf{y}) \geq \tau_0 \right\}, \tag{2.2}$$

where $R > 0$, $\tau_0 > 0$ are hyperparameters to be discussed later.

The likelihood (1.2) can be factorised into two components (see the Appendix for more details)

$$p(\mathbf{y}; \boldsymbol{\beta}, \tau) = p(\mathbf{y} | \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}); \boldsymbol{\beta}, \tau) \cdot p(\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}); \boldsymbol{\beta}, \tau), \tag{2.3}$$

where the second term in (2.3) is the probability density of the maximum likelihood estimates given parameters $\boldsymbol{\beta}$ and τ, and the first term is the likelihood of the data, conditional on the data yielding the given maximum likelihood estimates, i.e.,

$$p(\mathbf{y} | \hat{\boldsymbol{\beta}}, \hat{\tau}; \boldsymbol{\beta}, \tau) = \frac{p(\mathbf{y}; \boldsymbol{\beta}, \tau)}{\int_{T(\hat{\boldsymbol{\beta}}, \hat{\tau})} p(\mathbf{x}; \boldsymbol{\beta}, \tau) d\mathbf{x}}, \quad \mathbf{y} \in T(\hat{\boldsymbol{\beta}}, \hat{\tau}) \tag{2.4}$$

where

$$T(\hat{\boldsymbol{\beta}}, \hat{\tau}) = \left\{ \mathbf{x} \in \mathbb{R}^n : \hat{\boldsymbol{\beta}}(\mathbf{x}) = \hat{\boldsymbol{\beta}}, \hat{\tau}(\mathbf{x}) = \hat{\tau} \right\} \tag{2.5}$$

is the set of all data strings that yield maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}$. Additionally, due to the fact that for the Gaussian-linear regresson model the maximum likelihood estimates are sufficient statistics we also have the decomposition

$$p(\mathbf{y}; \boldsymbol{\beta}, \tau) = h(\mathbf{y}) \cdot p(\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}); \boldsymbol{\beta}, \tau), \tag{2.6}$$

where $h(\mathbf{y})$ is a normalization term that does not depend on $\boldsymbol{\beta}$ or τ. Comparing (2.3) and (2.6) we see that in this case

$$\begin{aligned} h(\mathbf{y}) &= p(\mathbf{y} | \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}); \boldsymbol{\beta}, \tau), \\ &\equiv p(\mathbf{y} | \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y})), \end{aligned} \tag{2.7}$$

which shows that the conditional density is independent of both $\boldsymbol{\beta}$ and τ. By the statistical independence of the maximum likelihood estimates $\hat{\boldsymbol{\beta}}(\mathbf{y})$ and $\hat{\tau}(\mathbf{y})$ we have

$$p_{\boldsymbol{\beta},\tau}(\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}); \boldsymbol{\beta}, \tau) = p_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}(\mathbf{y}); \boldsymbol{\beta}, \tau) \cdot p_{\tau}(\hat{\tau}(\mathbf{y}); \tau), \tag{2.8}$$

where $p_{\boldsymbol{\beta}}(\cdot)$ is the distribution of $\hat{\boldsymbol{\beta}}(\mathbf{y})$ given parameters $\boldsymbol{\beta}$ and τ and $p_{\tau}(\cdot)$ is the distribution of $\hat{\tau}(\mathbf{y})$ given parameter τ. The maximum likelihood estimates $\hat{\boldsymbol{\beta}}(\mathbf{y})$ are distributed as per a k-dimensional multivariate normal distribution

$$\hat{\boldsymbol{\beta}}(\mathbf{y}) \,|\, \boldsymbol{\beta}, \tau \sim N_k \left(\boldsymbol{\beta}, \tau \boldsymbol{\Sigma}^{-1} \right), \tag{2.9}$$

and (n/τ) times the maximum likelihood estimate $\hat{\tau}(\mathbf{y})$ is distributed as per a χ^2 distributed variable with $(n-k)$ degrees of freedom, i.e.,

$$p_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}(\mathbf{y}); \boldsymbol{\beta}, \tau) = \left(\frac{1}{2\pi\tau} \right)^{\frac{k}{2}} \cdot |\boldsymbol{\Sigma}|^{\frac{1}{2}} \cdot \exp\left(- \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\mathbf{y}))' \boldsymbol{\Sigma} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\mathbf{y}))}{2\tau} \right), \tag{2.10}$$

$$p_{\tau}(\hat{\tau}(\mathbf{y}); \tau) = \left(\frac{1}{2^{(n-k)/2} \, \Gamma\left(\frac{n-k}{2}\right)} \right) \cdot \left(\frac{n}{\tau} \right)^{\frac{n-k}{2}} [\hat{\tau}(\mathbf{y})]^{\left(\frac{n-k}{2} - 1\right)} \exp\left(- \frac{n\hat{\tau}(\mathbf{y})}{2\tau} \right). \tag{2.11}$$

Using the fact (see the Appendix for more details) that

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}(\mathbf{y}))'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}(\mathbf{y})) + n\hat{\tau}(\mathbf{y})$$

we see that the normalisation function $h(\cdot)$ is given by

$$h(\mathbf{y}) = \left(\frac{1}{2\pi}\right)^{\frac{n-k}{2}} \cdot |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot \left(\frac{1}{n}\right)^{\frac{n-k}{2}} [\hat{\tau}(\mathbf{y})]^{\left(1-\frac{n-k}{2}\right)} 2^{(n-k)/2}\, \Gamma\left(\frac{n-k}{2}\right). \tag{2.12}$$

We are now in a position to evaluate the parametric complexity term for the linear regression model, conditional on the hyperparameters R and τ_0 that define the region of integration $Y_1(R,\tau_0)$.

Using (1.3), (1.4), (2.10), (2.11) in (2.8) yields

$$\begin{aligned} p_{\boldsymbol{\beta},\tau}(\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y});\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y})) &= \left(\frac{n}{2}\right)^{\frac{n-k}{2}} \cdot \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \cdot \left(\frac{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}{\Gamma\left(\frac{n-k}{2}\right) e^{n/2}}\right) \cdot \left(\frac{1}{\hat{\tau}(\mathbf{y})}\right)^{\frac{k}{2}+1}, \\ &= A_{n,k} \cdot \left(\frac{1}{\hat{\tau}(\mathbf{y})}\right)^{\frac{k}{2}+1}. \end{aligned} \tag{2.13}$$

Using the fact that the maximised likelihood can be written as

$$p(\mathbf{y};\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y})) = h(\mathbf{y}) \cdot p_{\boldsymbol{\beta},\tau}(\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y});\hat{\boldsymbol{\beta}}(\mathbf{y}),\hat{\tau}(\mathbf{y})),$$

we can now write the denominator of the right-hand-side of (1.5), which we call $C_1(R,\tau_0)$, as the triple integral

$$C_1(R,\tau_0) = A_{n,k} \cdot \int_{\tau_0}^{\infty} \hat{\tau}^{-k/2-1} \int_{B(R)} \int_{T(\hat{\boldsymbol{\beta}},\hat{\tau})} h(\mathbf{y})\, d\mathbf{y}\, d\hat{\boldsymbol{\beta}}\, d\hat{\tau}. \tag{2.14}$$

where

$$B(R) = \{\boldsymbol{\beta} : \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} \leq nR\} \tag{2.15}$$

is the set of all regression models with fitted sum-of-squares less than nR, which is a k-dimensional hyperellipse defined by R and $\boldsymbol{\Sigma}$, and $T(\cdot)$ is given by (2.5). This triple integral representation works by first integrating over all the possible maximum likelihood estimate pairs, $(\hat{\boldsymbol{\beta}},\hat{\tau})$, that are implied by our set of data strings, $Y_1(R,\tau_0)$, and then for *each* of those pairs, integrating over the set of data strings, $T(\hat{\boldsymbol{\beta}},\hat{\tau})$, that yield those particular maximum likelihood estimates. By the equivalency between $h(\mathbf{y})$ and the conditional distribution (2.7), the inner integral in (2.14) is by definition equal to unity as it is taken over the set of strings for which the maximum likelihood estimates are constant. The remaining integrals are independent and we may evaluate the overall integral as a product of integrals

$$C_1(R,\tau_0) = A_{n,k} \cdot \int_{B(R)} d\hat{\boldsymbol{\beta}} \cdot \int_{\tau_0}^{\infty} \hat{\tau}^{-k/2-1}\, d\hat{\tau}.$$

The integral over $B(R)$ evaluates to the volume of a k-dimensional hyper-ellipse defined by $\boldsymbol{\Sigma}$ and R; we therefore have

$$\int_{B(R)} d\hat{\boldsymbol{\beta}} = V_k R^{\frac{k}{2}} \tag{2.16}$$

$$\int_{\tau_0}^{\infty} \hat{\tau}^{-k/2-1}\, d\hat{\tau} = \left(\frac{2}{k}\right)\left(\frac{1}{\tau_0}\right)^{\frac{k}{2}} \tag{2.17}$$

with

$$V_k = \frac{2(\pi n)^{\frac{k}{2}}}{k\,\Gamma(k/2)\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}},$$

where we have used the identity $\Gamma(a+1) = a\Gamma(a)$. Using (2.16) and (2.17) in (2.14) we have

$$C_1(R, \tau_0) = \left(\frac{2A_{n,k}V_k}{k}\right)\left(\frac{R}{\tau_0}\right)^{\frac{k}{2}}.$$

This leads to a complete stochastic complexity, conditional on the hyperparameters R and τ_0, of

$$\begin{aligned} -\log \bar{p}(\mathbf{y}; R, \tau_0) &= -\log\left[\frac{p(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}))}{C_1(R, \tau_0)}\right], \\ &= -\log\left[\left(\frac{1}{2\pi\hat{\tau}(\mathbf{y})e}\right)^{\frac{n}{2}} \cdot \left(\frac{k}{2A_{n,k}V_k}\right) \cdot \left(\frac{\tau_0}{R}\right)^{\frac{k}{2}}\right], \\ &= \frac{n}{2}\log\hat{\tau}(\mathbf{y}) + \frac{k}{2}\log\frac{R}{\tau_0} - \log\Gamma\left(\frac{k}{2}\right) - \log\Gamma\left(\frac{n-k}{2}\right) + \log\frac{4}{k^2} + \frac{n}{2}\log(\pi n). \end{aligned} \tag{2.18}$$

The stochastic complexity given by (2.18) is called the *first-level universal model* because we have formed a universal model for data arising from a Gaussian-linear regression model that does not depend on the $(k+1)$ parameters $\boldsymbol{\beta}$ or τ. However, it is clear that the criterion does depend crucially on the two hyperparameters R and τ_0 that are used to define the region of integration $Y_1(R, \tau_0)$. The removal of the dependency on these parameters is discussed in the next section.

3. Second-Level Universal Model

The most obvious way to "remove" the hyperparameters R and τ_0 is to set them to constant values that both the coder and decoder of the compressed data are aware of *a priori*. Unfortunately, as these two hyperparameters enter the codelength (2.18) through the term $(k/2)\log(R/\tau_0)$ the chosen values will have a crucial effect on the behaviour of the codelength when used as a model selection tool, i.e., to select which covariates to include in the model. An approach that was used in several earlier papers studying NML regression criteria was to set the hyperparameters to the values that minimise the codelength, say, $\hat{R}(\mathbf{y})$ and $\hat{\tau}_0(\mathbf{y})$. However, as noted by Rissanen [1], this means that the data will be coded using the distribution $\bar{p}(\mathbf{y}; \hat{R}(\mathbf{y}), \hat{\tau}_0(\mathbf{y}))$, yielding a message that is only decodable by someone who already knows the data. The Bayesian solution to this problem would be to put a prior density, say $w(\cdot)$ over the hyperparameters, but this would yield a density of the form

$$\bar{p}(\mathbf{y}; \hat{R}(\mathbf{y}), \hat{\tau}_0(\mathbf{y})) \cdot w(\hat{R}(\mathbf{y}), \hat{\tau}_0(\mathbf{y})),$$

which, to quote Rissanen, "is not quite right", as it is a density for the triple $(\mathbf{y}, \hat{R}, \hat{\tau}_0)$ rather than simply for $\mathbf{y}$. Rissanen suggests an ingenious solution to this problem, in which the first-level universal model is treated as a "likelihood" for the data, conditional on parameters R and τ_0, and the normalized maximum likelihood procedure is used to construct a (essentially) parameter free second-level universal model, i.e.,

$$\bar{p}(\mathbf{y}; R_1, R_2, \tau_1, \tau_2) = \frac{\bar{p}(\mathbf{y}; \hat{R}(\mathbf{y}), \hat{\tau}_0(\mathbf{y}))}{\int_{Y_2(R_1,R_2,\tau_1,\tau_2)} \bar{p}(\mathbf{x}; \hat{R}(\mathbf{x}), \hat{\tau}_0(\mathbf{x}))\, d\mathbf{x}}, \tag{3.1}$$

where the region of integration is

$$Y_2(R_1, R_2, \tau_1, \tau_2) = \left\{\mathbf{y} : R_1 \leq \hat{R}(\mathbf{y}) \leq R_2, \tau_1 \leq \hat{\tau}_0(\mathbf{y}) \leq \tau_2\right\}. \tag{3.2}$$

The maximum likelihood estimates of the hyperparameters R and τ_0 are relatively straightforward to derive. Recalling that both R and τ_0 define the region, $Y_1(R, \tau_0)$, of integration in the first-level universal model, we see from (2.18) that we wish to make R as small as possible, and τ_0 as large as

possible, subject to the resulting NML density being defined for the given data; that is, we require $\mathbf{y} \in Y_1(R, \tau_0)$. This leads to the following estimates

$$\begin{aligned} \hat{R}(\mathbf{y}) &= \frac{\hat{\boldsymbol{\beta}}'(\mathbf{y})\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}(\mathbf{y})}{n}, \\ \hat{\tau}_0(\mathbf{y}) &= \hat{\tau}(\mathbf{y}). \end{aligned}$$

An interesting property of (3.3) and (3.3) is that both of these estimates depend only on the maximum likelihood estimates for the likelihood $p(\mathbf{y}; \boldsymbol{\beta}, \tau)$. This means that, as in the case of the first-level universal model, we can take advantage of the fact that decomposition (2.3) and (2.6) are equivalent for the Gaussian linear regression model. The denominator of (3.1), denoted by $C_2(R_1, R_2, \tau_1, \tau_2)$, may then be written as

$$\begin{aligned} C_2(R_1, R_2, \tau_1, \tau_2) &= \int_{Y_2(R_1,R_2,\tau_1,\tau_2)} p(\mathbf{y}; \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y})) \cdot C_1(\hat{R}(\mathbf{y}), \hat{\tau}_0(\mathbf{y}))\, d\mathbf{y} \\ &= \int_{Y_2(R_1,R_2,\tau_1,\tau_2)} h(\mathbf{y}) \cdot p_{\boldsymbol{\beta},\tau}(\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}); \hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y})) \cdot C_1(\hat{R}(\mathbf{y}), \hat{\tau}_0(\mathbf{y}))\, d\mathbf{y} \\ &= \int_{Y_2(R_1,R_2,\tau_1,\tau_2)} h(\mathbf{y}) \cdot \left[A_{n,k} \left(\frac{1}{\hat{\tau}(\mathbf{y})} \right)^{\frac{k}{2}+1} \right] \cdot \left[\left(\frac{k}{2A_{n,k}V_k} \right) \left(\frac{\hat{\tau}_0(\mathbf{y})}{\hat{R}(\mathbf{y})} \right)^{\frac{k}{2}} \right] d\mathbf{y} \end{aligned}$$

We now make use of the fact that both $\hat{\tau}_0(\mathbf{y})$ and $\hat{R}(\mathbf{y})$ are functions of $\hat{\tau}(\mathbf{y})$ and $\hat{\boldsymbol{\beta}}(\mathbf{y})$, respectively. Defining $\hat{R}(\hat{\boldsymbol{\beta}}) = n^{-1}\hat{\boldsymbol{\beta}}'\boldsymbol{\Sigma}\hat{\boldsymbol{\beta}}$, and recalling that $\hat{\tau}_0(\mathbf{y}) = \hat{\tau}(\mathbf{y})$, we have

$$\begin{aligned} C_2(R_1, R_2, \tau_1, \tau_2) &= \left(\frac{k}{2V_k} \right) \cdot \int_{Y_2(R_1,R_2,\tau_1,\tau_2)} h(\mathbf{y}) \cdot \left(\frac{1}{\hat{\tau}(\mathbf{y})} \right) \cdot \left(\frac{1}{\hat{R}(\mathbf{y})} \right)^{\frac{k}{2}} d\mathbf{y}, \\ &= \left(\frac{k}{2V_k} \right) \cdot \int_{D(R_1,R_2)} \hat{R}(\hat{\boldsymbol{\beta}})^{-k/2} \cdot \int_{\tau_1}^{\tau_2} \hat{\tau}^{-1} \cdot \int_{T(\hat{\boldsymbol{\beta}},\hat{\tau})} h(\mathbf{y})\, d\mathbf{y}\, d\hat{\tau}\, d\hat{\boldsymbol{\beta}} \end{aligned} \tag{3.3}$$

where

$$D(R_1, R_2) = \{\boldsymbol{\beta} : R_1 \leq n\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} \leq R_2\}$$

is the set of all regression models with fitted-sum-of-squares between nR_1 and nR_2, and the set $T(\hat{\boldsymbol{\beta}}, \hat{\tau})$ is given by (2.5). Due to the equivalency between $h(\mathbf{y})$ and $p(\mathbf{y}|\hat{\boldsymbol{\beta}}(\mathbf{y}), \hat{\tau}(\mathbf{y}))$, the innermost integral, which is taken over the set of data strings for which the maximum likelihood estimates are fixed, yields unity by definition. We can then rewrite (3.3) as the product of integrals

$$C_2(R_1, R_2, \tau_1, \tau_2) = \left(\frac{k}{2V_k} \right) \cdot \int_{D(R_1,R_2)} \hat{R}(\hat{\boldsymbol{\beta}})^{-k/2}\, d\hat{\boldsymbol{\beta}} \cdot \int_{\tau_1}^{\tau_2} \hat{\tau}^{-1}\, d\hat{\tau}. \tag{3.4}$$

The integral with respect to $\hat{\tau}$ is straightforward:

$$\int_{\tau_1}^{\tau_2} \hat{\tau}^{-1}\, d\hat{\tau} = \log\left(\frac{\tau_2}{\tau_1} \right). \tag{3.5}$$

This leaves only the integral with respect to $\hat{\boldsymbol{\beta}}$. To evaluate this, we note the function $\hat{R}(\hat{\boldsymbol{\beta}})$ is constant over all coefficients $\hat{\boldsymbol{\beta}}$ that lie on the surface of the ellipsoids that are defined by $B(\hat{R})$. The set $D(R_1, R_2)$ is the union of surfaces of all the ellipsoids $B(\hat{R})$ for $\hat{R} \in (R_1, R_2)$. This implies that rather than integrate directly over $D(R_1, R_2)$, we can we integrate a suitable function of $\hat{R}$ over (R_1, R_2) with respect to the appropriate volume element, which in this case is simply the surface area of the ellipse defined by $B(\hat{R})$. The idea here is that we multiply each of the unique values of $\hat{R}^{-k/2}$ in (R_1, R_2) by

the "number" of coefficients $\hat{\boldsymbol{\beta}}$ such that $\hat{R}(\hat{\boldsymbol{\beta}}) = \hat{R}$ (which is the surface area of $B(\hat{R})$), and integrate this with respect to $\hat{R}$ over (R_1, R_2). The surface area is given by (see the Appendix for more details)

$$\text{Surf}(B(R)) = \left(\frac{kV_k R^{k/2-1}}{2}\right)$$

so that we have

$$\begin{aligned} \int_{D(R_1,R_2)} \hat{R}(\hat{\boldsymbol{\beta}})^{-k/2}\, d\hat{\boldsymbol{\beta}} &= \int_{R_1}^{R_2} \hat{R}^{-k/2} \cdot \left(\frac{kV_k \hat{R}^{k/2-1}}{2}\right) d\hat{R} \\ &= \left(\frac{kV_k}{2}\right) \log\left(\frac{R_2}{R_1}\right). \end{aligned} \tag{3.6}$$

Using (3.5) and (3.6) in (3.4) yields

$$C_2(R_1, R_2, \tau_1, \tau_2) = \left(\frac{k^2}{4}\right) \log\left(\frac{R_2\tau_2}{R_1\tau_1}\right).$$

We are now in a position to define the stochastic complexity of the string $\mathbf{y}$, conditional on the hyperparameters $(R_1, R_2, \tau_1, \tau_2)$:

$$\begin{aligned} -\log \bar{p}(\mathbf{y}; R_1, R_2, \tau_1, \tau_2) &= -\log\left[\frac{\bar{p}(\mathbf{y}; \hat{R}(\mathbf{y}), \hat{\tau}_0(\mathbf{y}))}{C_2(R_1, R_2, \tau_1, \tau_2)}\right], \\ &= \left(\frac{n-k}{2}\right)\log\hat{\tau}(\mathbf{y}) + \frac{k}{2}\log\hat{R}(\mathbf{y}) - \log\Gamma\left(\frac{k}{2}\right) - \log\Gamma\left(\frac{n-k}{2}\right) \\ &\quad + \frac{n}{2}\log(\pi n) + \log\log\left(\frac{R_2\tau_2}{R_1\tau_1}\right). \end{aligned} \tag{3.7}$$

We call (3.7) the codelength for the *second-level universal*. At first glance, it seems that by renormalization, we have succeeded only in swapping the two hyperparameters R and τ_0 for the four hyperparameters $(R_1, R_2, \tau_1, \tau_2)$. However, if we examine (3.7) we see that the renormalization has resulted in a very important outcomes: the four hyperparameters enter the second-level universal codelength formula through a term that is independent of k, $\hat{\tau}(\mathbf{y})$ or $\hat{\boldsymbol{\beta}}(\mathbf{y})$. Using the codelength formula for model selection involves comparing the codelengths of the competing models (with different combinations/numbers of covariates) and selecting those covariates that result in the shortest codelength. In this case, the addition of a term that is constant with respect to the choice of covariates implies that the choice of the hyperparameters will not affect the model selection behaviour of the second-level universal codelength (3.7). Thus, their choice is irrelevant and the term may be disregarded if we restrict ourselves to selecting only between competing linear regression models.

4. Appendix

4.1. **Least Squares Estimates.** The maximum likelihood estimates of the regression coefficients $\hat{\boldsymbol{\beta}}(\mathbf{y})$ are straightforward to derive. Begin by differentiating the negative log-likelihood with respect to $\boldsymbol{\beta}$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left\{\frac{n}{2}\log(2\pi\tau) + \left(\frac{1}{2\tau}\right)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\} = -\left(\frac{1}{\tau}\right)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{X}.$$

We then set the derivatives to zero and solve for $\boldsymbol{\beta}$. Defininining $\mathbf{0}_k$ as the column vector of zeros of size k, we have

$$-\left(\frac{1}{\tau}\right)\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{X}\right]' = \mathbf{0}_k,$$

Define $\mathbf{\Sigma} = \mathbf{X}'\mathbf{X}$; expanding the left hand term and multiplying both sides by τ yields

$$-\mathbf{X}'\mathbf{y} + \mathbf{\Sigma}\boldsymbol{\beta} = \mathbf{0}_k,$$

from which it is easy to arrive at the maximum likelihood estimator

$$\hat{\boldsymbol{\beta}}(\mathbf{y}) = \mathbf{\Sigma}^{-1}\mathbf{X}'\mathbf{y}.$$

To find the maximum likelihood estimator for τ, we note that the estimates $\hat{\boldsymbol{\beta}}(\mathbf{y})$ do not depend on τ and may be plugged directly into the negative log-likelihood. Ignoring terms that do not depend on τ, first expand the negative log-likelihood

$$\frac{n}{2}\log\tau + \left(\frac{1}{2\tau}\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{n}{2}\log\tau + \left(\frac{1}{2\tau}\right)(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{\Sigma}\boldsymbol{\beta})$$

Plugging $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\mathbf{y})$ into the right hand side of the above equation yields

$$\frac{n}{2}\log\tau + \left(\frac{1}{2\tau}\right)\left(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{\Sigma}^{-1}\mathbf{X}'\mathbf{y} + \mathbf{y}'\mathbf{X}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}\mathbf{X}'\mathbf{y}\right)$$

which, after cancellation simplifies to

$$\frac{n}{2}\log\tau + \left(\frac{1}{2\tau}\right)\left(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{y}\right).$$

Differentiating this expression with respect to τ and solving for zero yields the maximum likelihood estimate

$$\begin{aligned}\hat{\tau}(\mathbf{y}) &= \left(\frac{1}{n}\right)\left(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{y}\right), \\ &= \left(\frac{1}{n}\right)\mathbf{y}'\left(\mathbf{I}_n - \mathbf{X}'\mathbf{\Sigma}\mathbf{X}\right)\mathbf{y}.\end{aligned}$$

4.2. **Conditional Probability Decomposition.** Consider a probability distribution $p(\mathbf{y};\boldsymbol{\theta})$ for $\mathbf{y} \in Y \subset \mathbb{R}^n$, parameterised by $\boldsymbol{\theta}$. Let $f : Y \to Q$, with $Q \subseteq \mathbb{R}^q$, $q \leq n$, be a function from the dataspace to some other q-dimensional space. The probability distribution may be factorised in the following fashion

$$p(\mathbf{y};\boldsymbol{\theta}) = p(\mathbf{y}|f(\mathbf{y});\boldsymbol{\theta}) \cdot p(f(\mathbf{y});\boldsymbol{\theta}), \tag{4.1}$$

where $p(f(\mathbf{y});\boldsymbol{\theta})$ is the probability distribution for the function $f(\cdot)$ when data is generated by the distribution $p(\mathbf{y};\boldsymbol{\theta})$, and

$$p(\mathbf{y}|\bar{f};\boldsymbol{\theta}) = \frac{p(\mathbf{y};\boldsymbol{\theta})}{\int_{T(\bar{f})} p(\mathbf{y};\boldsymbol{\theta})d\boldsymbol{\theta}}, \quad \mathbf{y} \in T(\bar{f}),$$

with

$$T(\bar{f}) = \left\{\mathbf{x} : f(\mathbf{x}) = \bar{f}\right\}.$$

The set $T(\bar{f})$ is the set of all data strings $\mathbf{x}$ for which the function $f(\cdot)$ attains the same value. Thus, the decomposition (4.1) says the following: the probability of the data string $\mathbf{y}$, under the distribution $p(\mathbf{y};\boldsymbol{\theta})$, is equal to the probability of the function $f(\cdot)$ attaining the value $f(\mathbf{y})$ for data strings generated by $p(\mathbf{y};\boldsymbol{\theta})$, multiplied by the probability of the data string $\mathbf{y}$ arising from the model $p(\mathbf{y};\boldsymbol{\theta})$ when the value of the function $f(\cdot)$ is fixed at $f(\mathbf{y})$.

4.3. **Gaussian Linear Regression Identities.** We have the following equivalence

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\mathbf{y}))'\boldsymbol{\Sigma}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\mathbf{y})) + n\hat{\tau}(\mathbf{y}).$$

To see this, we note that

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}, \tag{4.2}$$

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\mathbf{y}))'\boldsymbol{\Sigma}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\mathbf{y})) = \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \mathbf{y}'\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{X}'\mathbf{y}, \tag{4.3}$$

$$n\hat{\tau}(\mathbf{y}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{X}'\mathbf{y}. \tag{4.4}$$

The result follows trivially by adding the right-hand-side of (4.3) to the right-hand-side of (4.4) to arrive at the right-hand-side of (4.2).

4.4. **Surface Area of Hyper-Ellipsoids.** Consider the k-dimensional ellipse defined by the set

$$E(nR, \boldsymbol{\Sigma}) = \left\{\mathbf{y} \in \mathbb{R}^k : \mathbf{y}'\boldsymbol{\Sigma}\mathbf{y} \leq nR\right\},$$

where $R > 0$ and $\boldsymbol{\Sigma}$ is positive definite. The volume of this ellipsoid is given by

$$\mathrm{Vol}\left(E(nR, \boldsymbol{\Sigma})\right) = \frac{2(\pi nR)^{\frac{k}{2}}}{k\,\Gamma(k/2)\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}}.$$

The surface area of the ellipsoid $E(nR, \boldsymbol{\Sigma})$ can be found by differentiating the volume with respect to the parameter R

$$\mathrm{Surf}\left(E(R, \boldsymbol{\Sigma})\right) = \frac{\partial}{\partial R}\left\{\mathrm{Vol}(E(R, \boldsymbol{\Sigma}))\right\} = \frac{(\pi n)^{\frac{k}{2}} R^{\frac{k}{2}-1}}{\Gamma(k/2)\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}}.$$

References

[1] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.
[2] ——, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.